计算机科学
COMPUTER SCIENCE

基于灰狼优化算法的信用评估样本均衡化与特征选择同步处理

储安琪，丁志军

引用本文

储安琪，丁志军. 基于灰狼优化算法的信用评估样本均衡化与特征选择同步处理[J]. 计算机科学, 2022, 49(4): 134-139.

CHU An-qi, DING Zhi-jun. Application of Gray Wolf Optimization Algorithm on Synchronous Processing of Sample Equalization and Feature Selection in Credit Evaluation[J]. Computer Science, 2022, 49(4): 134-139.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300075

基于灰狼优化算法的信用评估样本均衡化与特征选择同步处理

储安琪　丁志军
嵌入式系统与服务计算教育部重点实验室(同济大学)　上海 201804
上海市电子交易与信息服务协同创新中心(同济大学)　上海 201804
(1933013@tongji.edu.cn)

摘　要　随着互联网金融行业的迅速发展,面对海量数据,传统信用风险评估面临着挑战。信用评估中样本类别不均衡,且特征冗余度高,成为影响目前评估分类精度的关键因素。为了解决以上问题,提出了一种基于灰狼优化算法同步处理样本欠采样与特征选择的方法。该方法将分类器的性能作为灰狼优化算法的启发式信息,然后进行智能搜索,以得到最优样本与特征集的组合,并在原始灰狼算法中引入禁忌表策略,避免算法陷入局部最优。实验表明,该方法相较于其他方法有较大改进,在不同数据集上的表现均证明了该方法能够有效解决样本不均衡问题,降低特征空间维度,同时提高分类准确率。其在信用风险评估上相比原始数据准确率提高了3%左右,证实了该方法在信用评估领域的适用性与优越性。

关键词:信用评估;样本不均衡;特征选择;灰狼优化算法
中图法分类号　TP3-05

Application of Gray Wolf Optimization Algorithm on Synchronous Processing of Sample Equalization and Feature Selection in Credit Evaluation

CHU An-qi and DING Zhi-jun
Key Laboratory of Embedded System and Service Computing of Ministry of Education (Tongji University),Shanghai 201804,China
Shanghai Electronic Transactions and Information Service Collaborative Innovation Center (Tongji University),Shanghai 201804,China

Abstract　With the rapid development of Internet finance industry,traditional credit risk evaluation is facing challenges in the face of massive data. Due to the unbalanced sample categories and high feature redundancy in credit evaluation,it has become the key factor affecting the classification accuracy of current evaluation. In order to solve the above problems,a method based on gray wolf optimization algorithm is proposed to process the samples under sampling and feature selection synchronously. In this method,the performance of the classifier is taken as the heuristic information of the gray wolf optimization algorithm,and then the intelligent search is carried out to obtain the combination of the optimal sample and the feature set,and the tabu table strategy is introduced into the original gray wolf algorithm to avoid the algorithm falling into the local optimum. Experimental results show that the proposed method has a great improvement compared with other methods,and its performance on different data sets proves that it can effectively solve the problem of sample imbalance,reduce the dimension of feature space,and improve the accuracy of classification. Compared with the original data,the accuracy of credit risk evaluation is improved by about 3%,which proves the applicability and superiority of this method in the field of credit evaluation.

Keywords　Credit evaluation,Sample imbalance,Feature selection,Gray wolf optimization algorithm

1　引言

由于传统金融业的局限性,资本市场很难满足小额贷款借款人和投资者的借款需求。互联网金融具有准入门槛低、交易高效、贷款灵活方便的优势,使得这种交易模式在2007年开始在国内迅速传播[1]。在大数据时代背景下,互联网金融存在数据来源广、特征复杂化等特点,个人及企业的信用风险问题日益严重。信用风险指借款人由于种种原因,无法偿还贷款本息而导致贷款人损失的可能性。而信用评估作为管理信用风险的一个工具,对金融公司是否为个人或企业放贷起着决策性作用。在实际信贷业务中,可能发生违约的客户与正常客户相比数量较少,并且客户数据中通常包含多维特征,其中掺杂着许多无关的冗余特征,会严重干扰信用评估结果,因此对不平衡的信用数据进行均衡化处理以及特征

到稿日期:2021-03-08　返修日期:2021-07-14
基金项目:上海市科技创新行动计划(19511101300)
This work was supported by the Shanghai Science and Technology Innovation Action Plan(19511101300).
通信作者:丁志军(dingzj@tongji.edu.cn)

选择成为了研究者们广泛关注的问题。

在样本不均衡学习领域里，欠采样已被普遍用于缩小多数类和少数类之间的差距[2]，最简单的方法是随机欠采样，通过随机删除多数类中的某些样本，将剩余样本与少数类组成新的数据集。然而这种方法存在偶然性，容易造成多数类中的重要信息丢失，从而影响分类性能。为了克服这一缺点，Wilson 等提出了一种最近邻规则(Edited Nearest Neighbor，ENN)[3]，如果某样本类别与其最近的 3 个近邻样本中的两个或两个以上类别不同，则将它删除。Mani 等提出 Near Miss[4]来选择最接近少数类的多数类样本。周志华研究团队提出了 EasyEnsemble 算法和 BalanceCascade 算法[5]，EasyEnsemble 算法从多数类中随机抽取若干子集，然后将每个子集与少数类分别合并训练生成多个基分类器，最后组合多个分类器的结果；BalanceCascade 算法多次训练分类器，每次训练时删除那些分类正确的多数类样本。Lemaitre 等通过用 K-Means 算法的聚类质心替换多数类样本来进行欠采样[6]。还有研究运用深度神经网络来解决类极度不均衡问题[7]。

特征选择作为一种有效提高分类器性能的方法被广泛应用于风险管理领域。如 Fritz 等运用单变量线性判别法来衡量每个特征与信用水平的相关性程度[8]。Ding 等利用 F-Statistic 来衡量特征与被解释变量之间的相关性程度[9]。这类方法的运用主要包括前向选择法、后向剔除法以及逐步回归法，然而上述方法无法保证所选取的特征组合是趋于最优的。鉴于此，Hall 提出了基于相关性的特征选择方法(Correlation-based Feature Selection，CFS)[10]，既考虑了特征变量对目标判别的重要程度，还考虑了特征之间的冗余性。Tran 等用粒子群算法(Particle Swarm optimization，PSO)将特征离散化和特征选择合为一步来实现[11]，考虑了特征与特征之间的相关性。Zhang 等还提出了一种自然进化策略来进行特征选择[12]。

现有的研究中样本均衡化处理和特征选择大多是分开进行的，若样本均衡化在先，则会受到冗余特征的影响，使结果并非最优解，反之亦然。本文基于智能群体优化算法强大的寻优能力，提出一种封装式(Wrapper)方法，利用最新提出的且被证实具有明显优势的灰狼优化算法(Grey Wolf Optimizer，GWO)[13]同步筛选最优数据子集和特征子集，该算法在自然语言处理[14]、网络入侵检测[15]、汽车自动驾驶[16]和路径规划[17]等问题上均展示出不俗的表现。本文在原始灰狼算法中引入禁忌表策略以解决算法陷入局部最优问题，完成数据欠采样与特征选择工作，将其应用到信用评估中，通过实验验证了本文提出的方法的优越性。

2 相关理论

2.1 灰狼优化算法

灰狼优化算法是一种模拟狼群捕食过程中的智能行为而提出的新型群体优化算法。它的基本思想是根据社会结构将狼群分为 α 狼、β 狼、δ 狼和 ω 狼，其中 α 狼作为种群中的最高领导者起领头作用，β 狼和 δ 狼负责辅助 α 狼进行协作，ω 狼是种群中的最低等级，遵从其他领导者进行捕猎行动。

在整个优化过程中，α 狼始终是具有最优适应度的狼，即最优解记为 α，而 β 和 δ 被看作次优解，ω 狼根据 α，β 和 δ 3 个引导狼更新自己的位置信息。灰狼包围猎物移动的公式如下：

$$\vec{D}=|\vec{C}\cdot\vec{X}_p(t)-\vec{X}(t)| \tag{1}$$

$$\vec{X}(t+1)=\vec{X}_p(t)-\vec{A}\cdot\vec{D} \tag{2}$$

其中，$\vec{D}$ 表示当前狼个体与猎物之间的距离；t 为当前迭代次数；$\vec{A}$ 和 $\vec{C}$ 为系数向量；$\vec{X}_p$ 为猎物的位置向量；$\vec{X}$ 为灰狼的位置向量。$\vec{A}$ 和 $\vec{C}$ 的计算公式如下：

$$\vec{A}=2a\cdot\vec{r_1}-a \tag{3}$$

$$\vec{C}=2\vec{r_2} \tag{4}$$

$$a=2-\frac{2t}{t_{\max}} \tag{5}$$

其中，$\vec{r_1}$ 和 $\vec{r_2}$ 是[0，1]中的随机向量，a 是[0，2]上线性递减的系数。

其余灰狼 ω 根据 α，β 和 δ 3 个灰狼与自身所处位置信息进行更新，位置更新公式如下：

$$\begin{cases}\vec{D}_\alpha=|\vec{C}_1\cdot\vec{X}_\alpha(t)-\vec{X}(t)|\\ \vec{D}_\beta=|\vec{C}_2\cdot\vec{X}_\beta(t)-\vec{X}(t)|\\ \vec{D}_\delta=|\vec{C}_3\cdot\vec{X}_\delta(t)-\vec{X}(t)|\end{cases} \tag{6}$$

$$\begin{cases}\vec{X}_1=\vec{X}_\alpha(t)-\vec{A}_1\cdot\vec{D}_\alpha\\ \vec{X}_2=\vec{X}_\beta(t)-\vec{A}_2\cdot\vec{D}_\beta\\ \vec{X}_3=\vec{X}_\delta(t)-\vec{A}_3\cdot\vec{D}_\delta\end{cases} \tag{7}$$

$$\vec{X}(t+1)=\frac{\vec{X}_1+\vec{X}_2+\vec{X}_3}{3} \tag{8}$$

2.2 禁忌搜索

GWO 算法每次根据头狼的位置来更新所有种群个体位置，称为一次迭代，然后计算所有种群个体的适应度值，选取最优的 3 个个体，标记为 α，β，δ，其他个体进行位置更新，进入下一次迭代，直至满足收敛条件或者达到最大迭代次数，算法停止。算法每次迭代不断接近最优解，在初期算法具有较快的收敛速度，在迭代后期头狼若陷入局部最，优则其余种群个体根据决策层更新自己的位置时均会围绕在局部最优点附近，因此算法后期的全局搜索能力较差，容易陷入局部最，优从而影响最终结果，因此引入禁忌搜索使算法脱离局部最优。禁忌搜索时会使用禁忌表存放已经得到的局部最优解或优化过程，从而在进一步迭代中避开它，以此指导下一步的搜索方向，达到跳出局部最优的目的。

禁忌灰狼优化算法(Tabu Grey Wolf Optimizer，TGWO)在灰狼优化算法中引入禁忌搜索，其算法流程如下。

Step1 初始化种群参数，在搜索过程中根据式(3)—

式(5)计算参数 a,$\vec{A}$,$\vec{C}$。

Step2 计算每个灰狼个体的适应度值,从中选取适应度值最优的 3 个解作为第一代决策层 α,β,δ。

Step3 根据式(6)－式(8)计算其余灰狼与决策层引导狼的距离信息来更新其余灰狼的位置,再比较每个灰狼的适应度值,重新选取适应度值前 3 的解作为新一代 α,β,δ,以 α,β,δ 为当前解。

Step4 在当前解的设定邻域里随机选取候选解集合。

Step5 判断当前候选解是否存在于禁忌表中,若存在,则跳过此次迭代。

Step6 计算候选解的适应度值,选取其中适应度值最优的一个解作为最优候选解来替换当前解;如果当前最优候选解的适应度值优于决策层的优解,则替换它。

Step7 更新禁忌表,将最优候选解加入禁忌表,替换最早进入禁忌表的解。

Step8 若当前达到最大禁忌迭代次数,则跳转至 Step9,否则跳转至 Step4。

Step9 判断是否达到灰狼优化算法最大迭代次数,若满足,则跳转至 Step10,否则跳转至 Step3 进行下一次迭代。

Step10 输出最优解。

2.3 TGWO 算法分析

根据算法步骤对 TGWO 算法进行复杂度分析,这里计算每一步的运算次数并讨论算法的时间复杂性。Step1 在 D 维搜索空间下初始化 N 个个体需要 ND 次运算;*Step2* 计算每个灰狼个体的适应度值需要 N 次运算,根据参数训练分类模型,不同的分类器有着不同的时间复杂度,因此这里使用一般适应度函数的复杂度 O(D)来分析,从中选出适应度值最优的 3 个解最多需要 $3N-3$ 次运算,提取最优解的运算次数加 1;Step3 计算其余灰狼与决策层距离信息需要 $3(N-3)$次运算,距离函数复杂度为 O(D),重新更新选取适应度值前 3 的解并提取最优解需要 $3D-1$ 次运算;Step4 和 Step5 随机选取 M 个候选解以及判断其是否存在于禁忌表中的复杂度均为常数时间;Step6 计算候选解的适应度值需要 M 次运算,选出最优解并与当前最优解进行比较,最多需要 M 次运算;Step7－Step10 更新禁忌表并判断算法是否达到最大禁忌次数和最大迭代次数,其复杂度均为常数时间。算法禁忌迭代最多执行 t_1 次,TGWO 迭代最多执行 t_2 次,则经过近似和简化运算,TGWO 算法的时间复杂度约为:

$$O(ND)+t_2[O(ND)+t_1O(MD)]\approx O[t_2D(N+t_1M)]$$

3 样本不均衡处理和特征选择同步处理

3.1 方法流程

本文所提方法是将样本不均衡处理和特征选择两个过程合并进行。由于样本以及特征之间的组合与预测类之间的关系是非线性的[18],利用 TGWO 算法可以基于分类器结果实现数据和特征的组合优化,该方法的流程如图 1 所示。

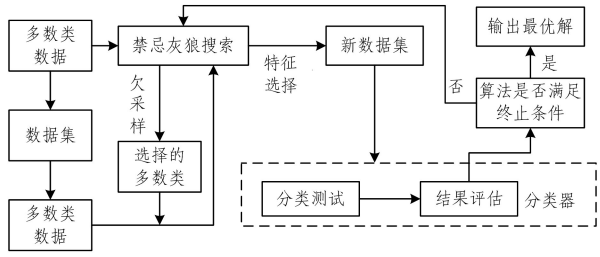


图 1 TGWO 算法的欠采样与特征选择流程图

Fig. 1 Flow chart of undersampling and feature selection based on TGWO algorithm

该方法从多数类数据中随机选择与少数类个数相等的样本,与少数类数据结合创建新的数据集,同时进行特征选择,使新数据集的特征空间得到降维。TGWO 算法中每个灰狼代表了样本与特征的组合,用其删除原始数据集中未选中的特征,同时挑选出多数类样本,与其余少数类样本一起放入分类器得到适应度值,通过比较选择出最优组合。

3.2 初始化种群

由于标准的 GWO 算法用于求解连续变量问题,不适用于样本与特征选择这类典型的离散空间组合优化,因此本文使用二进制编码对种群位置进行操作。设种群中灰狼数量为 m,数据集样本数量为 d_1,特征数量为 d_2,标志向量维度 $d=d_1+d_2$,通过随机方式对灰狼个体进行初始化,初始化公式如下:

$$X_i=[X_i^j],1\leqslant i\leqslant m,1\leqslant j\leqslant d \tag{9}$$

$$X_i^j=\begin{cases}0,\text{if } rand<0.5\\1,\text{if } rand\geqslant 0.5\end{cases},1\leqslant i\leqslant m,1\leqslant j\leqslant d \tag{10}$$

此时灰狼位置各维度在“0”和“1”之间转换,“1”表示选择该样本或特征,“0”表示不选择,如图 2 所示。

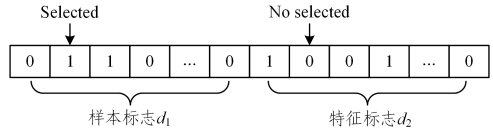


图 2 标志向量

Fig. 2 Label vector

灰狼个体位置更新时,通过映射函数 Sigmoid(见式(11))将移动后的连续位置向量映射到{0,1}上:

$$S(X_i^j(t))=\frac{1}{1+e^{(X_i^j(t))}} \tag{11}$$

$$X_i^j(t)=\begin{cases}0, & \text{if } rand<S(X_i^j(t))\\1, & \text{if } rand\geqslant S(X_i^j(t))\end{cases} \tag{12}$$

其中,X_i^j 表示第 i 个灰狼第 j 维的值,$rand$ 是[0,1]之间的随机数。这样种群更新后的连续空间位置经过映射函数离散为二进制状态,更新后的标志向量代表了新的组合样本和特征子集。

3.3 适应度函数

适应度函数是用来评判种群中每个个体的好坏。针对信用评估这个典型的分类问题,选用分类错误率(Error Rate)作为适应度函数的值,错误率最小时得到最优解,其具体定义为:

$$Fitness=Error\ Rate=\frac{FN+FP}{P+N} \tag{13}$$

其中，FN 为错误分类的正例数目，FP 为错误分类的负例数目，P 和 N 为所有的正例、负例数目。

4 实验分析

4.1 评估指标

分类模型的混淆矩阵如表 1 所列。

表 1 混淆矩阵

Table 1 Confusion matrix

	Actual true	Actual false
Predict true	*TP*	*FP*
Predict false	*FN*	*TN*

基于混淆矩阵的概念，我们可以定义以下指标。

准确率：所有正确分类个数占数据集总数的比率。

$$ACC=\frac{TP+TN}{TP+FN+TN+FP} \tag{14}$$

灵敏度：正确分类的正例占所有正例的比率。

$$SEN=\frac{TP}{TP+FN} \tag{15}$$

特异度：正确分类的负例占所有负例的比率。

$$SPE=\frac{TN}{FP+TN} \tag{16}$$

以特异度为横坐标、灵敏度为纵坐标，可以得到 ROC (Receiver Operating Characteristic Curve) 曲线，该曲线下的面积是 AUC(Area Under Curve)值，用它作为评估指标可以衡量模型的整体分类性能。AUC 取值为(0.5，1]，模型分类性能越好，该值就越大[19]。

4.2 对比实验

为了验证本文方法的有效性，选用著名的 UCI 机器学习数据库中的 6 个数据集进行实验，将所提方法与经典的欠采样及特征选择分步相结合的方法进行对比，对比方法包括 RUS(Random Under Sampling)、Near Miss 欠采样方法和 CFS、RFE(Recursive Feature Elimination)特征选择方法。其中，RFS 是一种基于模型的嵌入式(Embedded)特征选择方法，通过训练模型得到特征重要度，删除得分较低的特征。数据集描述如表 2 所列。

表 2 实验数据集描述

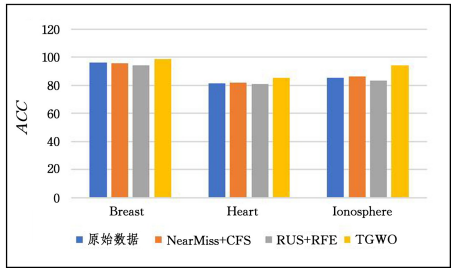
Table 2 Description of experimental data sets

数据集	样本数	多数类	少数类	特征数
Breast	699	458	241	9
Heart	270	150	120	13
Ionosphere	351	225	126	34
Pima	768	500	268	8
Sonar	208	111	97	60
Hepatitis	155	123	32	20

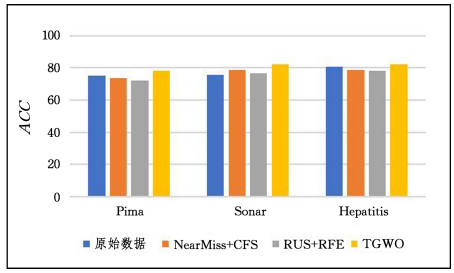
实验电脑的基本配置：处理器为 2.3 GHz Intel Core i7，内存 16 GB，操作系统为 Windows10 64 位，编程软件为 Matlab 2016a。

本文实验均使用支持向量机(Support Vector Machine，SVM)作为分类器，采用默认参数的线性核 LibSVM[20]，适应度函数同 3.3 节。为了测试算法的准确性，采用 5 折交叉验证的方法得到评估结果，即将数据集分成 5 份，轮流取其中 4 份作为训练数据、其余 1 份作为测试数据进行实验，5 次实验的平均值则为最终结果。本文算法的参数设置：种群规模 $N=30$，最大迭代次数 $t=30$，禁忌表长度为 6，邻域集合大小为 5。

为了便于比较，将 6 个数据集在不同方法上的评估结果绘制成直方图。从图 3 和图 4 可以看出，本文方法的分类准确率和 AUC 均高于用经典方法进行欠采样和特征选择处理的数据分类结果，由于 RUS 和 Near Miss 欠采样时容易丢失某些具备关键信息的多数类样本，导致准确率可能反而低于原始数据的结果。而且经典方法是基于经验公式拟合，考虑的影响因素较少，拟合结果多受数据分布影响。由此可以说明，本文方法不仅能将数据处理的两个关键步骤合二为一来降低复杂度，同时还能保证分类性能的提升。



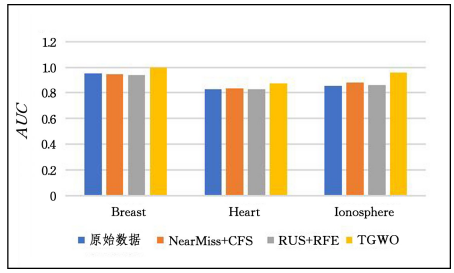
(a)准确率对比



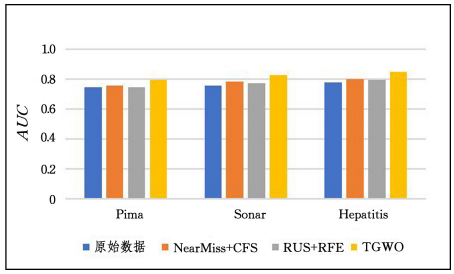
(b)准确率对比

图 3 数据集在不同方法处理后的准确率对比

Fig. 3 Accuracy comparison of data sets processed by different methods



(a)AUC 对比



(b)AUC 对比

图 4 数据集在不同方法处理后的 AUC 对比

Fig. 4 AUC comparison of data sets processed by different methods

在信用评估的应用上，实验选取了常用的两个公开信用数据集(German 和 Australian)，同样下载自 UCI 机器学习数据库，研究目的是对银行信贷申请客户是否会违约进行预测。澳大利亚信用数据集由 690 个样本构成，好客户有 383 个，坏客户有 307 个，每个样本包含 14 个特征属性。德国信用数据集由 1000 个样本构成，好客户有 700 个，坏客户有 300 个，每个样本包含 20 个特征属性，例如年龄、性别、工作情况、账户余额、借款期限等变量指标。由于涉及隐私，所有的属性名称都用相应的符号表示。

图 5 为用 TGWO 初始化 20 个种群个体、同步筛选样本与特征子集迭代 30 次的算法收敛图。明显可以看出，该算法在适应度值上呈现出了良好的收敛趋势，适用于信用评估领域。

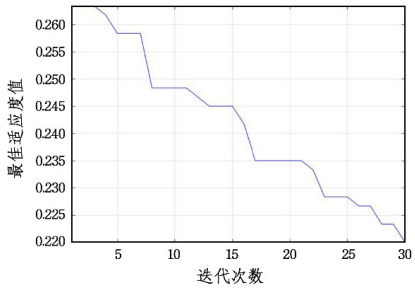


图 5 算法收敛图

Fig. 5 Algorithm convergence graph

将本文方法与仅使用 TGWO 进行单目标操作以及将 TGWO 替换成原始的 GWO 的方法进行比较，以此验证本文方法的合理性和性能。参数设置：种群规模 $N=50$，最大迭代次数 $t=100$，禁忌表长度为 6，邻域集合大小为 5。

表 3 和表 4 列出了本文方法处理过的数据与原始数据、TGWO 仅欠采样、TGWO 仅特征选择、GWO 同步处理过的数据在 SVM 上分类的结果，每种算法的评估结果均取 10 次独立实验的平均值。从表 3 可知，在澳大利亚信用数据集上，本文方法的分类准确率为 87.36%，*AUC* 为 0.914；从表 4 可知，在德国信用数据集上，本文方法的分类准确率为78.98%，*AUC* 为 0.801，均为最高值。结果表明，与原始数据直接分类对比，TGWO 在欠采样和特征选择上都得到了良好的应用成果。然而很多时候违约样本占极少数，模型无法对其充分训练学习。由表 4 明显可以看出，仅进行特征选择虽然整体准确率有所提高，但特异度较低，少数类样本分类结果不佳，而信用评估的主要目的是尽量找出可能违约的客户以降低风险，特异度低会影响对违约客户的判断；仅进行欠采样提高了特异度，但没有降低特征空间维度，准确率也没有特征选择后的数据结果好，易将好客户判别为坏客户。与仅用 TGWO 单独欠采样和单独特征选择的单目标方法对比证明了本文方法的有效性，并且样本均衡化与特征选择同步处理既能减少样本数量、平衡数据集，又能使特征维度降低，使训练效率明显提高。与 GWO 的对比结果验证了 TGWO 的寻优能力高于原始的种群优化算法。

表 3 澳大利亚数据集上算法的对比结果

Table 3 Comparison result of algorithms on Australian dataset

	准确率/%	灵敏度/%	特异度/%	*AUC*
原始数据	84.44	83.28	86.56	0.887
TGWO(欠采样)	84.78	83.16	86.82	0.896
TGWO(特征选择)	86.38	85.13	87.68	0.904
GWO	86.95	85.37	88.48	0.911
TGWO	**87.36**	**86.12**	**89.37**	**0.914**

表 4 德国数据集上算法的对比结果

Table 4 Comparison results of algorithms on German dataset

	准确率/%	灵敏度/%	特异度/%	*AUC*
原始数据	75.35	81.65	51.65	0.722
TGWO(欠采样)	76.33	70.02	80.67	0.748
TGWO(特征选择)	77.19	83.81	53.57	0.732
GWO	78.21	80.49	76.35	0.790
TGWO	**78.98**	**81.24**	**77.58**	**0.801**

结束语 本文提出了一种同步处理样本均衡化与特征选择的方法，该方法通过一种改进的灰狼优化算法 TGWO 实现，然后将其应用于信用评估。实验表明，在信用评估中，所提方法在分类准确率和 AUC 两个评估指标方面均具有明显优势，能大幅度提高信用评估性能。未来将进一步思考研究如何解决算法搜索空间大、如何提高算法效率等问题，并考虑该方法在其他领域中的应用。

参 考 文 献

[1] SUN H, WANG B. Research on Credit Risk Assessment of Online Network Credit Based on GBDT[C]// 2020 International Conference on Big Data in Management. 2020.

[2] PENG M, ZHANG Q, XING X, et al. Trainable Undersampling for Class-Imbalance Learning[C]// AAAI Conference on Artificial Intelligence. 2019.

[3] WILSON D L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data[J]. IEEE Transactions on Systems Man & Cybernetics, 1972, SMC-2(3): 408-421.

[4] MANI I, ZHANG J. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction[C]// ICML Workshop on Learning from Imbalanced Datasets. 2003.

[5] LIU X Y, WU J, ZHOU Z H. Exploratory Undersampling for Class-Imbalance Learning[J]. IEEE Transactions on Cybernetics, 2009, 39(2): 539-550.

[6] LEMAITRE G, NOGUEIRA F, ARIDAS C K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning[J]. Journal of Machine Learning Research, 2016, 18(17): 1-5.

[7] LIU Y, YANG K. Credit Fraud Detection for Extremely Imbalanced Data Based on Ensembled Deep Learning[J]. Journal of Computer Research and Development, 2021, 58(3): 539-547.

[8] FRITZ S, HOSEMANN D. Restructuring the credit process: behaviour scoring for German corporates[J]. Intelligent Systems in Accounting Finance & Management, 2000, 9(1): 9-21.

[9] DING C, PENG H. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics

and Computational Biology,2005,3(2):185-206.

[10] HALL M. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning[C]// Proceedings of the 17th International Conference on Machine Learning. San Francisco, Morgan Kaufmann,2000:359-366.

[11] TRAN B,XUE B,ZHANG M. A New Representation in PSO for Discretization-Based Feature Selection[J]. IEEE Transactions on Cybernetics,2018,48(6):1733-1746.

[12] ZHANG X,LI Z S. Research on Feature Selection Algorithm Based on Natural Evolution Strategy[J]. Journal of Software, 2020,31(12):3733-3752.

[13] MIRJALILI S,MIRJALILI S M,LEWIS A. Grey Wolf Optimizer[J]. Advances in Engineering Software,2014,69:46-61.

[14] ZHANG P Y,HUANG X Z,LI M Z,et al. Hybridization between Neural Computing and Nature-Inspired Algorithms for a Sentence Similarity Model Based on the Attention Mechanism [J]. ACM Transactions on Asian and Low-Resource Language Information Processing,2021,20(1):1-21.

[15] MISHRA S,DWIVEDULA R,KSHIRSAGAR V,et al. Robust Detection of Network Intrusion using Tree-based Convolutional Neural Networks[C]// 8th ACM IKDD CODS and 26th COMAD. 2021.

[16] INDU S,SRIVASTAVA S,SHARMA V. Optimal Camera Placement and Orientation of A Multi-camera System for Self Driving Cars[C]// Proceedings of the 2020 4th International Conference on Vision,Image and Signal Processing. 2020:1-5.

[17] LIU J,CHEN Z,ZHANG Y,et al. Path Planning of Mobile Robots based on Improved Genetic Algorithm[C]// 2020 2nd International Conference on Robotics,Intelligent Control and Artificial Intelligence. 2020.

[18] WANG W J,SUN Y Y,SUN H L,et al. Research on Multi-source Heterogeneous Data Classification Based on Multi-objective Optimization Technology[J]. Computer and Digital Engineering,2020,48(1):130-136.

[19] ZHOU M. Credit Evaluation for Hybrid Grey Wolf Optimization and Least Squares Support Vector Machine Approach[J]. Journal of Chengdu University of Technology(Science & Technology Edition),2019,46(4):507-512.

[20] CHANG C C,LIN C J. LIBSVM:A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27.

CHU An-qi, born in 1995, postgraduate. Her main research interests include data mining and machine learning.

DING Zhi-jun, born in 1974, Ph.D, Professor, Ph. D supervisor, is a senior member of China Computer Federation. His main research interests include intelligent software engineering, cloud computing and services, big data credit reporting and financial risk control.

(责任编辑:柯颖)